ROBOT AUDITORY SYSTEM BASED ON CIRCULAR MICROPHONE ARRAY FOR HOME SERVICE ROBOTS

Keun-Chang Kwak
Dept. of Control, Instrumentation, and Robot Engineering, Chosun University
375 Seosuk-dong Dong-gu, Gwangju, 501-759, Korea

Keywords: Robot auditory system, Circular microphone array, Home service robots, Sound source localization, Speaker recognition, Speech enhancement.

Abstract: In this paper, we develop robot auditory system including speaker recognition, sound source localization, and speech enhancement based on circular microphone array for home service robots. These techniques are concerned with audio-based Human-Robot Interaction (HRI) that can naturally interact between human and robot through audio information obtained from microphone array and multi-channel sound board. The robot platform used in this study is wever, which is a network-based intelligent home service robot. The experimental results show the effectiveness of the presented audio-based HRI components from the constructed speaker and sound localization database.

1 INTRODUCTION

During the past few years, we have witnessed a rapid growth in the number and variety of applications of robots, ranging from conventional industrial robots to intelligent service robots. Conventional industrial robots perform jobs and simple tasks by following pre-programmed instructions for humans in factories. On the other hand, the main objective of the intelligent service robot is to adapt for the necessities of life as accessibility to human life increases. While industrial robots have been widely used in many manufacturing industries, intelligent service robots are still in elementary standard. Although the intelligent robots have been brought to public attention, the development of intelligent service robots remains as a matter to be researched further. Recently, there has been a renewal of interest in Human-Robot Interaction (HRI) for intelligent service robots. Among various HRI components, we especially focus on audio-based HRI including speech enhancement, speech recognition, speaker recognition, sound source localization, sound source separation, and gender/age classification. We shall deal with some of audio-based HRI components. The robot platform used in this paper is wever, which is a network-based intelligent home service robot equipped with multi-channel sound board and three low-cost condenser microphones. Finally, we shall show the performance of the developed techniques such as speaker recognition, sound localization, and speech enhancement among audio-based HRI components from the databases constructed in u-robot test bed.

2 AUDIO-BASED HRI

In this section, we present text-independent speaker recognition based on MFCC (Mel-Frequency Cepstral Coefficients) and GMM (Gaussian Mixture Model), sound source localization based on ESI (Excitation Source Information), and speech enhancement based on PBF (Phase-error Based Filter) with circular microphone array equipped with intelligent service robot.

2.1 Speaker Recognition

Firstly the EPD (Endpoint Detection) algorithm is performed to analyze speech signal obtained from speaker. Here the speech signal is detected by log energy and zero crossing. After detecting signal, the feature extraction step is performed by six stages to obtain MFCC. These stages consist of pre-emphasis, frame blocking, hamming window, FFT (Fast Fourier Transform), triangular bandpass filter, and

Kwak K. (2009).
ROBOT AUDITORY SYSTEM BASED ON CIRCULAR MICROPHONE ARRAY FOR HOME SERVICE ROBOTS.
In *Proceedings of the 6th International Conference on Informatics in Control, Automation and Robotics - Robotics and Automation*, pages 475-478
DOI: 10.5220/0002248504750478

cosine transform. For simplicity, we use 11 MFCC parameters except for the first order. In what follows, we construct GMM (Reynolds and Rose, 1995) frequently used in conjunction with text-independent speaker recognition to represent speaker's individual model in robot environments. Figure 1 shows the signal detected by log energy and zero crossing rate. Figure 2 shows the block diagram for feature extraction.

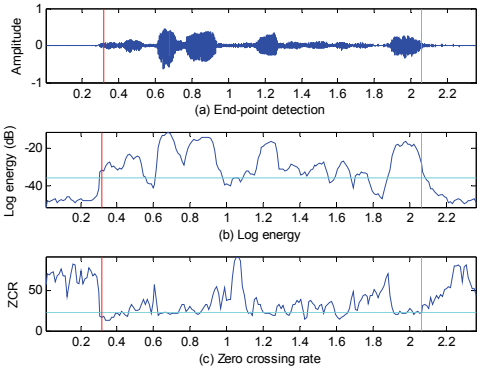

Figure 1: Endpoint detection.

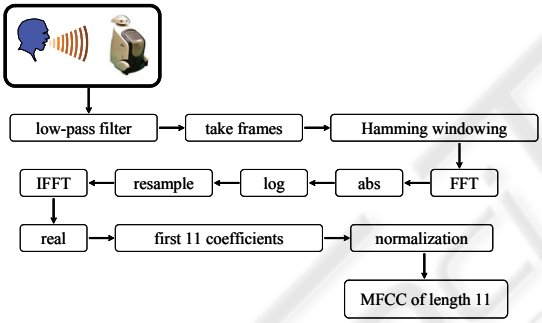

Figure 2: Block diagram for feature extraction.

2.2 Sound Source Localization

Sound source localization is performed by excitation source information and reliable angel estimation to determine the time-delay between each two microphones from speech source when robot's name is called. The time-delay based on excitation source information is comprised on two main stages. The first stage is to estimate time-delay from speech signals collected by three microphones. For this, the segmented signals by the endpoint detection should be detected. Here the speech signal of robot's name used in this study is wever. In order to perform time-delay estimation based on excitation source information, we firstly need to obtain linear prediction residual. This error includes the important information about excitation source during speech production. The linear prediction residual has a large value around the instants of glottal closure for voiced speech. However, these residuals should be transformed to derive critical information from short segments of linear prediction residual due to large fluctuations in amplitude. In the second stage, the values of linear prediction residual are transformed by computing the Hilbert envelop of linear prediction residual signal (Raykar and Yegnanarayana, 2005). Figure 3 shows linear prediction residual and Hilbert envelope. Figure 4 shows time-delay estimation from transformed signals obtained by Hilbert envelope.

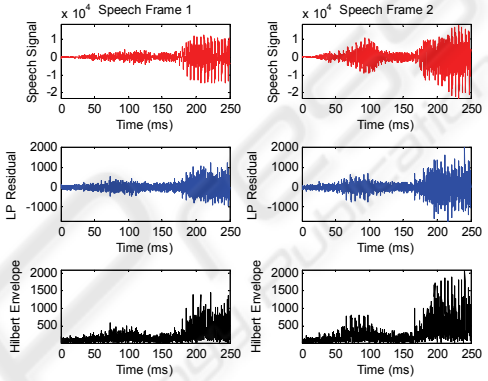

Figure 3: Linear prediction residual and Hilbert envelope.

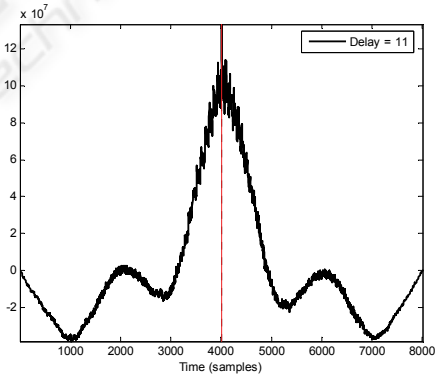

Figure 4: Time-delay estimation.

2.3 Speech Enhancement

This section present multi-microphone signal processing for speech recognition based on phase-error based filtering (Aarabi and Shi, 2004). This filtering performs time-frequency masking in the STFT (Short-time Fourier Transform) domain. For each pair of input frames, their phase-error spectrum is computed and used to modulate the amplitude spectrum. High error yields lower masking values. This has the effect of reducing time-alignment mismatch for each frequency bin, which is supposed

to be related to reverberation and noise. Therefore, this method involves obtaining time-varying, or alternatively, time-frequency, phase-error filters based on prior knowledge regarding the time difference of arrival of the speech source of interest and the phase of the signals recorded by the microphones. Figure 5 shows the signals obtained from three microphones and enhanced signal, respectively.

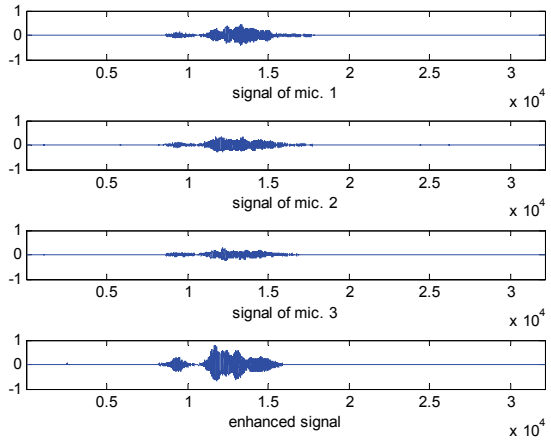

Figure 5: Speech enhancement by three microphones.

3 EXPERIMENTAL RESULTS

In this section, we use speaker database to evaluation the performance of the presented speaker recognition system (Kwak et al., 2007). Figure 6 shows "wever" equipped with three microphones and sound boards shown in Figure 7. The database is constructed by audio recording of 20 speakers. The data set consists of 30 sentences for each speaker and channel. For simplicity, we use only single microphone and 2200 sentences. The recording was done in u-robot test bed. The audio is stored as a mono, 16bit, 16kHz, and WAV file. The experimental results within 3 meter showed a good recognition performance of 94.5% recognition rate. However, the recognition performance at 4 and 5 meter showed 87.5% and 83%, respectively (see Figure 8).

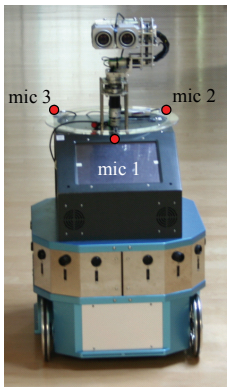

Figure 6: Robot platform-"wever".

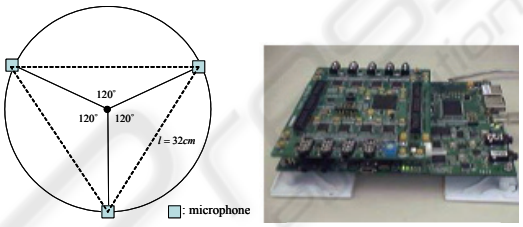

Figure 7: Arrangement of microphones and sound board.

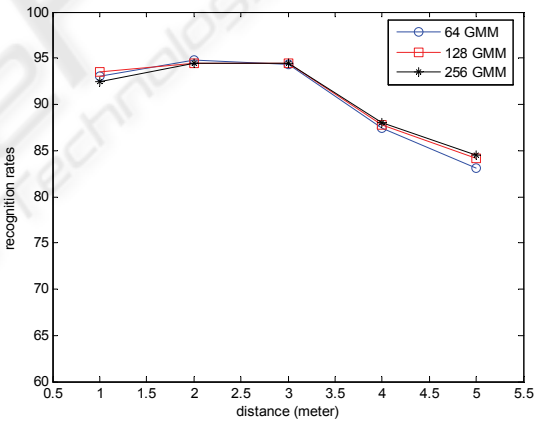

Figure 8: Recognition performance.

On the other hand, the localization success rate is considered as performance measure for sound localization. The localization success rate is computed by FOV (Field of View) of robot camera because sound localization is used with face detection when robot moves toward caller. The database used in this study was constructed in u-robot test bed environment that is similar with home environment to evaluate the sound localization algorithm (Kwak et al., 2008). The data set (M1 and M2) consists of 72 speeches at each meter from 1 meter to 3 meter. The localization success rate of the presented method is 92.3%. The presented method showed a better localization performance (about

20%) in comparison to that of TDOA and GCC-PHAT (see Figure 9 and 10).

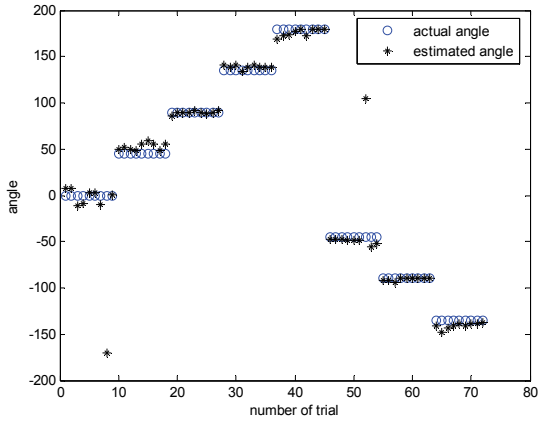

Figure 9: Actual angles and estimated angles (M1 set).

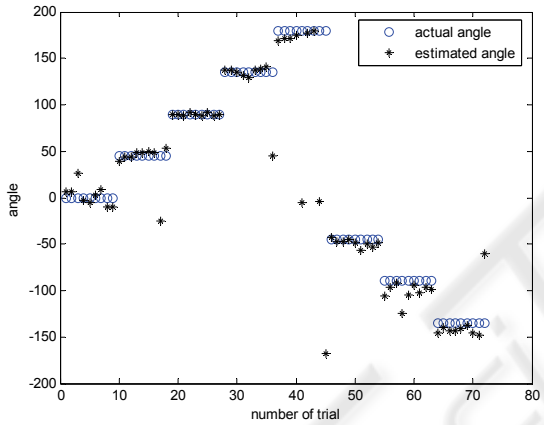

Figure 10: Actual angles and estimated angles (M2 set).

4 CONCLUSIONS

We have developed some of audio-based HRI components for intelligent home service robots. These components are composed of speaker recognition based on MFCC-GMM, sound source localization based on ESI and reliable angle estimation, and speech enhancement based on PBF. We have showed the usefulness and effectiveness of the developed techniques through the performance obtained from the constructed databases. On the basis of these components, we shall continuously develop other techniques such as sound source separation, gender/age classification, and fusion of information obtained from multi-microphones for humanlike robot auditory system.

REFERENCES

Reynolds, D. A., Rose, R. C., 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, pp. 72-83.

Kwak, K. C., Kim, H. J., Bae, K. S., Yoon, H. S., 2007. Speaker identification and verification for intelligent service robots. In *International Conference on Artificial Intelligence (ICAI2007)*, Las Vegas, May.

Raykar, V. C., Yegnanarayana, B., Prasanna, S. R. M., Duraiswami R., 2005. Speaker localization using excitation source information in speech. *IEEE Trans. on Acoustic, Speech, and Signal Processing*, vol. 13, no. 5, pp. 751-761.

Kwak, K. C., Kim, S. S., 2008. Sound source localization with the aid of excitation source information in home robot environments. *IEEE Trans. on Consumer Electronics*, vol. 54, no. 2, pp. 852-856.

Aarabi, P., Shi, G., 2004. Phase-based dual-microphone robust speech enhancement. *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 34, no. 4, pp. 1763-1773.